QualityTrails: Data Quality Provenance as a Basis for Sensemaking

Christian Bors[1], Theresia Gschwandtner[1], Silvia Miksch[1], and Johannes Gärtner[2]
[1]Institute of Software Technology & Interactive Systems (ISIS), Vienna University of Technology, Austria, {*surname*}*@ifs.tuwien.ac.at*
[2]XIMES GmbH, Vienna, Austria, *gaertner@ximes.com*

ABSTRACT

Visual Analytics prototypes increasingly support human sensemaking through providing Provenance information. For data analysts the challenge of knowledge generation starts with assessing the quality of a data set, but Provenance is not yet utilized to aid this task. This position paper aims at characterizing the complexity of Visual Analytics methods introducing Provenance in Data Quality by highlighting the challenges of (1) generating Provenance from Data Quality Control and (2) sensemaking based on Data Quality Provenance.

Keywords: data provenance, analytic provenance, sensemaking, data quality, quality metrics, visual data analysis

Index Terms: Human-centered Computing [Visualization]: Visualization application domains—Visual Analytics Mathematics of computing [Probability and statistics]: Statistical paradigms—Exploratory data analysis

1 INTRODUCTION

In Data Quality (DQ) assessment one of the central questions is, 'Is the Data Quality good enough for analysis to produce meaningful results?' The quality of data analysis is highly dependent on the quality of the underlying data. Thus, a prerequisite of any data analysis, such as creating visualizations and performing analytical reasoning, is assessing and improving DQ. Data cleansing is an iterative task that requires user expertise and domain knowledge of the data provided [7]. DQ control can be understood as a combination of data quality assessment, the data cleansing process, as well as applying transformations to change a data set's structure. Kandel et al. [7] argue that integrating interactive and visual systems could facilitate these tasks as well as data verification.

Yet, the analyst is left with the decision about when quality is sufficient to start analysis, or if the data is worth further manipulating at all. Sensemaking is an integral part of Visual Analytics (VA). During DQ assessment the analyst needs to take into account not only the actual data, but also implicit information, like how the data was created or its transformation history. A data set already might have been analyzed by someone else, generating a transformation history or other insight. This information could be helpful for further analysis. Provenance conceives this information and makes it available to the data analyst. Establishing a model for sensemaking to grasp the context of a data set benefits knowledge discovery.

2 CHALLENGES

We reviewed the state-of-the-art of Provenance generation [11], Provenance in VA [10], as well as sensemaking in VA [1, 13], and lastly Provenance in DQ assessment [5, 2]. In the following sections we illustrate our results, i.e., the current VA approaches that combine DQ with Provenance to aid analysts in their task of making sense of data. Furthermore we derive open problems and challenges for Provenance in DQ analysis and contemplate possible solutions.

2.1 Provenance from Data Quality Control

Data Provenance information is primarily utilized for resolving conflicting data sets and estimating data reliability based on lineage [3, 14]. Hartig [5] suggests to use a Provenance model in DQ to assess metrics like accuracy, reliability, or timeliness of data, which partly conforms with the above mentioned use of conflict resolution.

However, there are just a few approaches that denote how Provenance information could be used for DQ improvement and assessment. In the following sections we propose approaches to outline which Provenance information is suited to aid these tasks and how it should be gathered.

Generating Provenance from Data Cleansing Operations. Some data analysis tools incorporate the concept of logging the actions of data manipulation [8]. Generating Provenance from cleansing operations is a promising approach. By now, this information is merely presented in textual form and used for tracking purposes rather than DQ assessment. Provenance information from data exploration and transformation can be obtained by tracing transformation steps, cleansing operations, etc. in a log for later inspection.

The Open Provenance Model [11] (OPM) has been developed to depict Provenance information through an *Artifact*, *Process*, and *Agent* model. DQ assessment and cleansing operations can be applied to this model as means of tracking the action history of quality assessment, employing *data sources*, similar to artifacts, *transformation functions*, comparable to processes, and *analysts*, interpretable as agents. The model's design is generic enough to support this task. It provides a good overview of which actions have been taken by other analysts. However, this approach does not consider implicit information about data generation sources or information based on the analyst's experiences while cleansing operations are omitted. Thus, it is necessary to either investigate the extensibility of this model or to find other solutions that are suited to convey this information.

Generating Provenance from Annotated Data. As a common way of propagating insightful information to collaborators or analysts annotations are employed to further analyze the data [9]. They can be seen as a type of Provenance information and allow for manually adding information about the conditions under which the data have been created or manipulated. This information is important to analysts in order to correctly assess the DQ and to be aware of all kinds of background information.

Hullman et al. [6] proposed automatically adding narrative annotations to line-charts of stock visualizations. They stress the importance of using annotations as an additional information source to support sensemaking. In existing data analysis approaches annotation is not directly incorporated, but analysts rely on informal information and consider it in their sensemaking process. We propose administrating annotations about data sources and quality cleansing operations as Provenance information.

Generating Provenance from Quality Metrics. In Data Quality Management one approach to measuring Data Quality is computing Data Quality Metrics [12] (QM). The aim is to find structural or measurement errors by means of computation. This is a task that requires comprehensive knowledge about the error sources and causes, as well as how they manifest in the data. Metrics can be used to both give overview on a data set and simultaneously give detailed information on specific values, by being calculated on multiple granularity levels. Errors in the data set are

propagated to high level overviews and can still be easily tracked by browsing lower aggregation layers.

With quality problems being resolved over time, also the quality measures improve and indicate a trend during DQ assessment. We propose utilizing development of the data quality – as indicated by QM computed at different points in time – as Provenance. We contemplate that an analyst can determine if the quality is sufficient for analysis from assessing gradual quality improvement over time, comparing the current status to the data's original condition.

We have described approaches to generate Provenance from data cleansing operations, from annotations, as well as from meta information based on QM. Logging this information allows their integration into computation processes and it can be used to deduce patterns and learn about domain specific traits. Another challenge is to design means that foster the integration of DQ Provenance into sensemaking.

2.2 Sensemaking based on Data Quality Provenance

It is not enough to capture Provenance information about DQ, it also needs to be integrated into sensemaking. Making sense of data is a highly complex task, which requires the analyst to be aware of the circumstances under which the data have been generated and by which contingencies they were influenced. The diversity of Provenance information can be significant. It is necessary to determine ways of efficiently presenting various types of Provenance information to the analyst without obstructing data cleansing operations.

In general, DQ improvement is used to prepare a data set for subsequent analysis. Attfield et al. [1] suggest that analysts aim at generating a model of sensemaking based on their semantic knowledge in combination with available information. We identify three iterative phases in the course of DQ assessment and sensemaking where the analyst combines his/her semantic knowledge with information about the data set and its respective Provenance information:
(1) The analyst decides if the data is usable, based on the Provenance information that has been provided.
(2) The analyst has a certain goal in mind what to do with the data in the subsequent analysis and thus he/she transforms and refines the data to achieve an output that supports sensemaking in this specific context.
(3) Based on the Provenance information the analyst determines his/her confidence in the data, and thus, in the analysis results and interprets the outcome accordingly.

One way of further supporting the sensemkaing process is the use of efficient visualizations, providing the necessary information in a suitable format.

Visualizing Provenance from Data Quality Assessment. Provenance for sensemaking in DQ has the potential to provide substantial additional information to the analyst. It is necessary to develop means of visually propagating this information to him/her. Analytic Provenance approaches resort to graph- or tree-like visualization techniques to develop visual representations of Provenance graphs [5, 13, 10]. Attfield et al. [1] suggest to employ visualization prototypes to provide indicators that let analysts hypothesize on the data.

Carata et al. [4] claim that little research has been put into alternative visualization techniques, aside from node-link representations. We propose novel ideas on how to utilize Provenance information to generate visual aids in a DQ assessment environment. Which types of visual aids are suited for this task depends, of course, on the type of information. QM measure data properties over time, and are usually normalized. This implies that a continuous multivariate line-chart could properly visualize such information and support the decision-making process of the data analyst. Manual annotations could serve as guiding-points in either data table views or in the suggested line-chart visualizations of QM development over time, similar to Hullman et al. [6].

We contemplate combining visualizations of different Provenance information types into interactive views that employ linking and brushing. Within these multiple views annotations could be used to accentuate significant events and draw conclusions. Providing such visualizations in addition to Provenance graphs would provide enriched means for DQ aware data analysis, i.e., different kinds of visualization for different analysis tasks.

3 Outlook

In our upcoming research we aim at tackling the challenges characterized above by developing a DQ control prototype that incorporates data cleansing and transformation operations as well as employing Provenance information to support analysts in their sensemaking tasks.

ACKNOWLEDGMENTS This work is part of the the Laura Bassi Centre of Expertise CVAST is funded by the Austrian Federal Ministry of Economy, Family and Youth (project number: 822746).

References

[1] S. J. Attfield, S. K. Hara, and B. L. W. Wong. Sensemaking in visual analytics: Processes and challenges. In J. Kohlhammer and D. Keim, editors, *EuroVAST 2010: Intern. Symp. on VAST*, pages 1–6, Bordeaux, France, 2010. Eurographics Association.

[2] C. Batini and M. Scannapieco. *Data Quality: Concepts, Methodologies and Techniques (Data-Centric Systems and Applications)*. Springer Verlag New York, Inc., Secaucus, NJ, USA, 2006.

[3] P. Buneman, S. Khanna, and W. C. Tan. Why and where: A characterization of data provenance. In J. V. d. Bussche and V. Vianu, editors, *Intern. Conf. DB Theory*, pages 316–330. Springer, LNCS 1973, 2001.

[4] L. Carata, S. Akoush, N. Balakrishnan, T. Bytheway, R. Sohan, M. Seltzer, and A. Hopper. A primer on provenance. *Queue*, 12(3):10:10–10:23, Mar. 2014.

[5] O. Hartig and J. Zhao. Using web data provenance for quality assessment. In J. Freire, P. Missier, and S. S. Sahoo, editors, *SWPM*, volume 526 of *CEUR Workshop Proceedings*. CEUR-WS.org, Oct. 2009.

[6] J. Hullman, N. Diakopoulos, and E. Adar. Contextifier: Automatic generation of annotated stock visualizations. In *Proc. SIGCHI Conf. Human Factors in Computing Systems*, CHI '13, pages 2707–2716, New York, NY, USA, 2013. ACM.

[7] S. Kandel, J. Heer, C. Plaisant, J. Kennedy, F. van Ham, N. H. Riche, C. Weaver, B. Lee, D. Brodbeck, and P. Buono. Research directions in data wrangling: Visualizations and transformations for usable and credible data. *Inf. Vis. Journal*, 10(4):271–288, 2011.

[8] S. Kandel, R. Parikh, A. Paepcke, J. M. Hellerstein, and J. Heer. Profiler: Integrated statistical analysis and visualization for data quality assessment. In *Proc. Intern. Working Conf. Advanced Visual Interfaces*, AVI '12, pages 547–554, New York, NY, USA, 2012. ACM.

[9] Q. Li, A. Labrinidis, and P. Chrysanthis. User-centric annotation management for biological data. In J. Freire, D. Koop, and L. Moreau, editors, *Provenance and Annotation of Data and Processes*, volume 5272 of *Lecture Notes in Computer Science*, pages 54–61. Springer Berlin Heidelberg, 2008.

[10] J. Lu, Z. Wen, S. Pan, and J. Lai. Analytic trails: Supporting provenance, collaboration, and reuse for visual data analysis by business users. In *Proc. 13th IFIP TC 13 Int. Conf. HCI - Vol. IV*, INTERACT'11, pages 256–273, Berlin, Heidelberg, 2011.

[11] L. Moreau, B. Clifford, J. Freire, J. Futrelle, Y. Gil, P. Groth, N. Kwasnikowska, S. Miles, P. Missier, J. Myers, B. Plale, Y. Simmhan, E. Stephan, and J. V. d. Bussche. The open provenance model core spec. (v1.1). *Future Gen. Computer Systems*, 27(6):743 – 756, 2011.

[12] S. Sadiq, editor. *Handbook of Data Quality*. Springer Verlag, Berlin, Heidelberg, 2013.

[13] Y. B. Shrinivasan and J. J. van Wijk. Supporting the analytical reasoning process in information visualization. In *Proc. SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 1237–1246, New York, NY, USA, 2008. ACM.

[14] Y. L. Simmhan, B. Plale, and D. Gannon. A survey of data provenance in e-science. *SIGMOD Rec.*, 34(3):31–36, Sept. 2005.